Katahira *et al. EURASIP Journal on Advances in Signal Processing* (2016) 2016:11
DOI 10.1186/s13634-015-0301-3

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer

Hiroki Katahira[1], Nobutaka Ono[2,3], Shigeki Miyabe[1], Takeshi Yamada[1] and Shoji Makino[1*]

Abstract

In this paper, we propose a new microphone array signal processing technique, which increases the number of microphones virtually by generating extra signal channels from real microphone signals. Microphone array signal processing methods such as speech enhancement are effective for improving the quality of various speech applications such as speech recognition and voice communication systems. However, the performance of speech enhancement and other signal processing methods depends on the number of microphones. Thus, special equipment such as a multichannel A/D converter or a microphone array is needed to achieve high processing performance. Therefore, our aim was to establish a technique for improving the performance of array signal processing with a small number of microphones and, in particular, to increase the number of channels virtually by synthesizing *virtual microphone* signals, or extra signal channels, from two channels of microphone signals. Each virtual microphone signal is generated by interpolating a short-time Fourier transform (STFT) representation of the microphone signals. The phase and amplitude of the signal are interpolated individually. The phase is linearly interpolated on the basis of a sound propagation model, and the amplitude is nonlinearly interpolated on the basis of β divergence. We also performed speech enhancement experiments using a maximum signal-to-noise ratio (SNR) beamformer equipped with virtual microphones and evaluated the improvement in performance upon introducing virtual microphones.

Keywords: Microphone array signal processing, Speech enhancement, Virtual microphone, Maximum SNR beamformer, β divergence

1 Introduction

Speech processing applications, such as voice communication and speech recognition systems, have become more common in recent years. To realize high-performance applications, a technique is needed to reduce the noise or interference contained in microphone signals. Therefore, there have been many studies on noise reduction and speech enhancement involving the use of beamformers. One typical speech enhancement approach is microphone array signal processing, which uses spatial information obtained with multiple microphones [1]. However, the performance of many speech enhancement methods with microphone arrays depends on the number of microphones, and the performance may degrade when a small number of microphones are used. Recently, recording equipment with a small number of microphones, such as IC recorders and mobile phones, has become common. Therefore, speech signal processing techniques are expected to be widely employed to realize high-performance speech enhancement with few microphones or recording channels.

Underdetermined blind source separation (BSS) for a mixture of sources whose number exceeds the number of microphones has been widely studied as a typical framework for array signal processing with a small number of microphones [2].

In this problem, conventional linear array signal processing is ineffective because nontarget sources can only be canceled accurately by linear processing when there are fewer sources than microphones. One typical approach to underdetermined BSS is the statistical modeling of observations using latent variables [3, 4]. Using latent variables,

*Correspondence: maki@tara.tsukuba.ac.jp
[1] University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan
Full list of author information is available at the end of the article

the ill-posed estimation problem of underdetermined BSS can be effectively formulated as an optimization problem, and iterative optimization methods such as the EM algorithm can be used to estimate the parameters required for nonlinear signal processing for underdetermined BSS. By using the latent variables, we design time-frequency masks which reduce the nontarget sources and enhance the target source. However, the use of time-frequency masks leads to too much discontinuous zero padding of the extracted signals, and therefore, they tend to contain musical noise, which is undesirable for audio applications.

In this paper, to extend the general linear array signal processing framework to enable its direct applicability to underdetermined observations, we propose a method for increasing the number of channels using *virtual microphones*, which are extra virtual channels of recordings synthesized from two real microphones. By using the increased number of channels as the signal processing input, speech enhancement with a small number of microphones is improved. The virtual microphone signal is generated by interpolation in the short-time Fourier transform (STFT) domain, which is individually conducted for the phase angle and amplitude (absolute value). We interpolate the phase linearly, whereas for the amplitude, the interpolation is based on β divergence [5]. This method is applicable to various applications including speech enhancement, source separation, direction of arrival (DOA) estimation, and dereverberation by beamformers and has high versatility. To evaluate the effectiveness of the proposed method in speech enhancement, we apply the proposed method to a maximum signal-to-noise ratio (SNR) beamformer whose number of input channels is increased using virtual microphones.

Since several approaches have been proposed to increase the number of channels or to generate virtual microphones for similar or different purposes, below we summarize the relationship between our proposed method and conventional methods.

First, a similar approach to our method of introducing virtual microphones, namely, increasing the number of linear array signal processing channels, has been studied by several groups [6–10]. While our proposed method generates a virtual signal in the audio signal domain, the signal generation in conventional methods is carried out in the power domain [6, 7, 10] or in a higher-order statistical domain [8, 9]. Although changing the domain can improve the DOA estimation and speech enhancement performance, the processed signal suffers from heavy distortion.

Another method of generating a signal in the audio signal domain has been proposed in the field of spatial audio acquisition [11, 12]. In this method, the DOA of the source is estimated using two distributed microphone arrays. Given the estimated source positions and the signal measured at a real reference microphone, a virtual microphone signal is obtained by applying suitable gains to a reference signal. However, the generated virtual microphone signal cannot be employed as an extra observation to cancel more sources.

In general, if we increase the number of microphones, we have more spatial information and more freedom in controlling the spatial directivity pattern. As a result, we can achieve better performance. Therefore, if we can increase the number of microphones by approximately following an appropriate rule, we can expect better performance in linear array signal processing.

The structure of this paper is as follows. In Section 2, we state the problem. In Section 3, we explain and formulate the generation of virtual microphone signals. In Section 4, we explain the maximum SNR beamformer, which is a speech enhancement method. In Section 5, we experimentally evaluate the performance of the maximum SNR beamformer with virtual microphones. In Section 6, we present our conclusion.

2 Linear model for speech enhancement

In typical speech enhancement methods, a microphone signal is modeled using the following mixing model in the time-frequency domain. Let $s_i(\omega, t)$ be the ith source signal at an angular frequency ω in the tth frame, and let $x_j(\omega, t)$ be the microphone signal at the jth microphone. Signals can be modeled as

$$\begin{aligned}\mathbf{x}(\omega, t) &= \left[x_1(\omega, t), \cdots, x_M(\omega, t) \right]^{\mathrm{T}} \\ &\approx \sum_{i=1}^{N} \mathbf{a}_i(\omega) s_i(\omega, t), \end{aligned} \tag{1}$$

$$\mathbf{a}_i(\omega) = \left[a_{1,i}(\omega), \cdots, a_{M,i}(\omega) \right]^{\mathrm{T}}, \tag{2}$$

where $a_{j,i}(\omega)$ is the transfer function from the ith source to the jth microphone and $\{\cdot\}^{\mathrm{T}}$ stands for the transposition of a matrix. Speech enhancement by beamforming is conducted by constructing a multichannel filter given by

$$\mathbf{w}(\omega) = \left[w_1(\omega), \cdots, w_M(\omega) \right]^{\mathrm{T}} \tag{3}$$

to reduce the nontarget sources or background noise from microphone signal $\mathbf{x}(\omega, t)$ and enhance the target speech, where $w_n^*(\omega)$ is the filter for the nth channel and $\{\cdot\}^*$ denotes complex conjugation. The enhanced signal $y(\omega, t)$ is given as

$$y(\omega, t) = \mathbf{w}^{\mathrm{H}}(\omega) \mathbf{x}(\omega, t), \tag{4}$$

where $\{\cdot\}^{\mathrm{H}}$ stands for the conjugate transposition of a matrix.

The nontarget sources can be reduced when $\mathbf{w}(\omega)$ is orthogonal to all the transfer functions $\mathbf{a}_i(\omega)$ of the nontarget sources as follows:

$$\mathbf{w}^{\mathrm{H}}(\omega)\mathbf{a}_i(\omega) = 0, \quad \forall i \neq i_{\mathrm{T}}, \tag{5}$$

$$\begin{aligned} y(\omega,t) &= \mathbf{w}^{\mathrm{H}}(\omega)\sum_{i=1}^{N}\mathbf{a}_i(\omega)s_i(\omega,t) \\ &= \mathbf{w}^{\mathrm{H}}(\omega)\mathbf{a}_{i_{\mathrm{T}}}(\omega)s_{i_{\mathrm{T}}}(\omega,t), \end{aligned} \tag{6}$$

where the i_{T}th source is the target. However, in an underdetermined case, when the dimension M of the microphone signal vectors is smaller than the number N of source signals, such a filter $\mathbf{w}(\omega)$ does not generally exist. The beamformer can reduce directional noise from only $(M-1)$ directions when M is the number of microphones. Therefore, a larger number M of microphones are needed to realize speech enhancement with high performance for a mixture consisting of the number N of sources.

3 Increasing the number of channels using virtual microphones

As described in the preceding section, it is preferable to use more microphones than sources to realize good speech enhancement performance. However, large-scale and costly equipment such as a microphone array or an A/D converter is required for recording with three or more channels. On the other hand, two-channel stereo recording is available in small, easily available equipment such as mobile phones and portable recorders. Therefore, we propose a method for increasing the number of recording channels virtually by generating additional signal channels.

3.1 Approach to virtual microphone signal generation

In our virtual microphone array technique, we create arbitrary channels of virtual microphone signals by using two channels of real microphone signals, then we perform array signal processing using microphone signals consisting of both real and virtual microphone signals as shown in Fig. 1. Virtual microphone signals are generated as estimates of signals at a virtual microphone placed at a point where there is no real microphone. A virtual microphone signal $v(\omega,t)$ is generated as an observation estimated by interpolation for a virtual microphone placed at a point with a distance ratio of $\alpha : (1-\alpha)$ from the positions of two real microphones (Fig. 2). Multiple collinear virtual microphones are generated by setting the virtual microphone interpolation parameter α to different values.

Virtual microphone signals must be generated by nonlinear processing because linearly independent signals are required to increase the number of channels for the signal processing input. Thus, we discuss appropriate nonlinear processing for generating a virtual microphone signal. As previously noted, a virtual microphone signal is generated as an estimate of a signal at a point where there is no real microphone; thus, we consider the relationship between the value of the virtual microphone interpolation parameter α and a wavefront propagating towards a microphone.

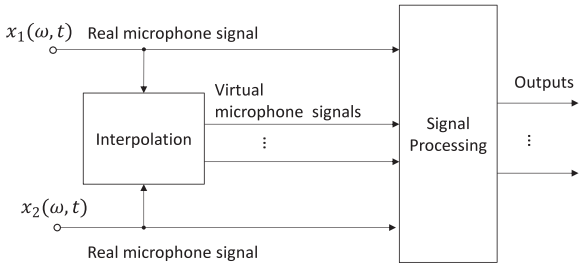

Fig. 1 Block diagram of signal processing using virtual microphone array technique

In speech enhancement, a microphone signal normally consists of a mixture of multiple sources. In this paper, we assume W-disjoint orthogonality (W-DO) [13, 14] for a mixed signal to simplify the observation model for the mixture. W-DO refers to the strong sparseness of a signal in the time-frequency domain, where it is assumed that a component from a single source dominates one time-frequency slot of a discrete STFT. In the proposed method, virtual microphone signals are generated in each time-frequency bin; thus, the relationship between the virtual microphone position and the propagating wavefront can be modeled as the propagation of a single wavefront.

3.2 Virtual microphone signal generation by interpolation

Here, we formulate the interpolation of the phase and amplitude separately. The phase and amplitude of the

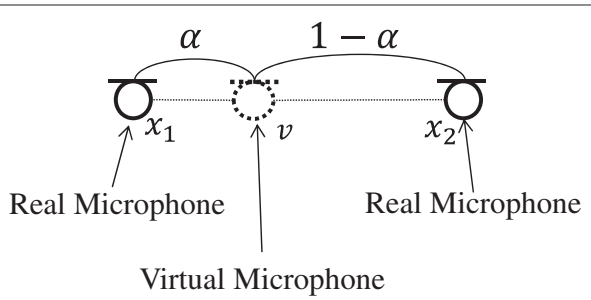

Fig. 2 Arrangement of real and virtual microphones

signal at microphone i are denoted by A_i and ϕ_i and are respectively given as

$$A_i = |x_i(\omega, t)|, \tag{7}$$

$$\phi_i = \angle x_i(\omega, t) = \arctan \frac{\mathrm{Im}(x_i(\omega, t))}{\mathrm{Re}(x_i(\omega, t))}. \tag{8}$$

We employ different models for phase and amplitude interpolation and simplify each formulation. The separate interpolation of the phase and amplitude introduces nonlinearity into the virtual signal generation, which is necessary to increase the number of channels.

3.2.1 Linear interpolation of phase

Assuming W-DO, as introduced in the previous section, we consider that a single wavefront propagates in each time-frequency bin. The propagating wave can be approximated as a plane wave when the acoustic wave arrives from a distance. Then, the phase of the signal has the following linear relationship with the spatial position:

$$\phi_\mathrm{v} = (1-\alpha)\phi_1 + \alpha\phi_2. \tag{9}$$

Note that the observed phase has an aliasing ambiguity given by $\phi_i \pm 2n_i\pi$ with integer n_i. In this phase interpolation, we assume that the microphone signal has no spatial aliasing with a sufficiently small microphone interval and that

$$|\phi_1 - \phi_2| \leq \pi. \tag{10}$$

3.2.2 Amplitude interpolation based on β divergence

As described in the previous section, the linear interpolation of the phase angle is based on the propagation model of a single planar wavefront. The signal amplitude must also be interpolated in accordance with an appropriate rule.

However, the physical modeling of the amplitude difference is not as simple as that of the phase difference because the amplitude depends on the distance between the source and the microphones in addition to the DOA. Thus, instead of interpolation based on some physical assumption, we utilize a distance measure in the interpolation. As an adjustable distance measure, we use β divergence.

β divergence is a widely used distance measure for nonnegative values such as amplitude. For instance, β divergence is used as the cost function for nonnegative matrix factorization (NMF) [5, 15]. β divergence is equivalent to Itakura-Saito divergence ($\beta = 0$), Kullback-Leibler divergence ($\beta = 1$), and Euclidean divergence ($\beta = 2$). Note that β divergence also corresponds to the far-field model ($\beta = 2$) and the near-field model. The β divergence between the signal amplitude of a virtual microphone A_v and that of the ith real microphone A_i is defined as

$$D_\beta(A_\mathrm{v}, A_i) = \begin{cases} A_\mathrm{v}\left(\log A_\mathrm{v} - \log A_i\right) + (A_i - A_\mathrm{v}) & (\beta = 1), \\ \frac{A_\mathrm{v}}{A_i} - \log\frac{A_\mathrm{v}}{A_i} - 1 & (\beta = 0), \\ \frac{A_\mathrm{v}^\beta}{\beta(\beta-1)} + \frac{A_i^\beta}{\beta} - \frac{A_\mathrm{v}A_i^{\beta-1}}{\beta-1} & (\text{otherwise}). \end{cases} \tag{11}$$

Note that D_β is continuous at $\beta = 0$ and $\beta = 1$. For β-divergence-based interpolation, we derive the amplitude A_v that minimizes the sum σ_β of the β divergence between the amplitude of a real microphones signal and a virtual microphone signal weighted by the virtual microphone interpolation parameter α,

$$\sigma_{D_\beta} = (1-\alpha) D_\beta(A_\mathrm{v}, A_1) + \alpha D_\beta(A_\mathrm{v}, A_2), \tag{12}$$

$$A_{\mathrm{v}\beta} = \mathrm{argmin}_{A_\mathrm{v}}\, \sigma_{D_\beta}. \tag{13}$$

Differentiating σ_{D_β} with respect to A_v and setting it to 0, the interpolated amplitude extended using β divergence is obtained as

$$A_{\mathrm{v}\beta} = \begin{cases} \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right) & (\beta = 1), \\ \left((1-\alpha)A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}). \end{cases} \tag{14}$$

Similar to β divergence D_β, $A_{\mathrm{v}\beta}$ is continuous at $\beta = 1$ because

$$\begin{aligned} A_{\mathrm{v}1} &= \lim_{\beta\to 1}\left((1-\alpha)A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} \\ &= \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right). \end{aligned} \tag{15}$$

When β is set to 1, the interpolated phase and amplitude are given as the following unified equation:

$$\nu = \exp\left((1-\alpha)\log x_1 + \alpha \log x_2\right), \tag{16}$$

and we call the interpolation based on Eq. (16) *complex logarithmic interpolation*. Additionally, the result of β-divergence-based interpolation is assumed to be the $\beta - 1$ norm of the vector $[(1-\alpha)x_1, \alpha x_2]^\mathrm{T}$, which is composed of the amplitude weighted by α. Therefore, taking the limits of $\beta \to +\infty$ and $\beta \to -\infty$, the interpolation corresponds to the selection of the following maximum and minimum values, respectively:

$$A_{\mathrm{v}\beta} = \begin{cases} \max(A_1, A_2) & (\beta \to +\infty), \\ \min(A_1, A_2) & (\beta \to -\infty). \end{cases} \tag{17}$$

In this paper, we substitute the β divergences at these infinite limits with the above maximum and minimum values. The virtual microphone signal is obtained as follows in terms of the interpolated phase and amplitude:

$$\nu = A_{\mathrm{v}\beta}\exp\left(j\phi_\mathrm{v}\right). \tag{18}$$

Note that the linear interpolation of the phase angle is defined in the domain of arbitrary real numbers α, not only in the range $0 \leq \alpha \leq 1$. On the other hand,

the β-divergence-based interpolation of the amplitude is defined only in the domain of $0 \leq \alpha \leq 1$ when β is set to $\beta \neq 1$. Therefore, in this paper, we deal solely with without extrapolation, where α is confined to $0 \leq \alpha \leq 1$.

4 Speech enhancement with maximum SNR beamformer

We apply the virtual microphone array technique to a maximum SNR beamformer [16] to evaluate its performance. A maximum SNR beamformer requires the covariance matrices of the target-active period and target-inactive period as prior information for speech enhancement.

The maximum SNR beamformer has the advantage of being available when the source direction is unknown because it requires no information about the sound direction such as its steering vectors. Therefore, it is easy to apply our virtual microphone to the maximum SNR beamformer.

4.1 Construction of maximum SNR beamformer

In a maximum SNR beamformer, the filter $\mathbf{w}(\omega)$ is designed to maximize the ratio $\lambda(\omega)$ of the power between the target-active period Θ_{T} and the target-inactive period Θ_{I}:

$$\lambda(\omega) = \frac{\mathbf{w}^{\mathrm{H}}(\omega)\mathbf{R}_{\mathrm{T}}(\omega)\mathbf{w}(\omega)}{\mathbf{w}^{\mathrm{H}}(\omega)\mathbf{R}_{\mathrm{I}}(\omega)\mathbf{w}(\omega)}, \tag{19}$$

where $\mathbf{R}_{\mathrm{T}}(\omega)$ and $\mathbf{R}_{\mathrm{I}}(\omega)$ represent the covariance matrices of the target-active period and target-inactive period, respectively. The covariance matrices are calculated as

$$\mathbf{R}_{\mathrm{T}}(\omega) = \frac{1}{|\Theta_{\mathrm{T}}|}\sum_{t\in\Theta_{\mathrm{T}}} \mathbf{x}_{\mathrm{T}}(\omega,t)\mathbf{x}_{\mathrm{T}}^{\mathrm{H}}(\omega,t), \tag{20}$$

$$\mathbf{R}_{\mathrm{I}}(\omega) = \frac{1}{|\Theta_{\mathrm{I}}|}\sum_{t\in\Theta_{\mathrm{I}}} \mathbf{x}_{\mathrm{I}}(\omega,t)\mathbf{x}_{\mathrm{I}}^{\mathrm{H}}(\omega,t), \tag{21}$$

where $\mathbf{x}_{\mathrm{T}}$ is the microphone signal vector in the target-active period and $\mathbf{x}_{\mathrm{I}}$ is the microphone signal vector in the target-inactive period. The filter $\mathbf{w}(\omega)$ that maximizes the ratio $\lambda(\omega)$ is given as the eigenvector corresponding to the maximum eigenvalue of the following generalized eigenvalue problem:

$$\mathbf{R}_{\mathrm{T}}(\omega)\mathbf{w}(\omega) = \lambda(\omega)\mathbf{R}_{\mathrm{I}}(\omega)\mathbf{w}(\omega). \tag{22}$$

4.2 Scaling compensation of beamformer

Since the maximum SNR beamformer $\mathbf{w}(\omega)$ has a scaling ambiguity, the beamformer is compensated similar to [17] as:

$$\mathbf{w}(\omega) \leftarrow b_k(\omega)\mathbf{w}(\omega), \tag{23}$$

where $b_k(\omega)$ is the kth component of $\mathbf{b}(\omega)$ given by

$$\mathbf{b}(\omega) = \frac{\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}(\omega)}{\mathbf{w}^{\mathrm{H}}(\omega)\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}(\omega)}, \tag{24}$$

$$\mathbf{R}_{\mathbf{x}}(\omega) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(\omega,t)\mathbf{x}^{\mathrm{H}}(\omega,t). \tag{25}$$

5 Speech enhancement experiments involving multiple directional sources

To evaluate the effectiveness of introducing virtual microphones, we conducted speech enhancement experiments using a maximum SNR beamformer.

5.1 Experimental conditions

The layout of the sources and real microphones is shown in Fig. 3, and the other experimental conditions are shown in Table 1. We used two samples of Japanese speech and one sample of English speech for the target signals, and we performed five DOA experiments for each sample target signal, giving a total of 15 trials for different combinations of the target DOA and speech samples. We used a mixture of eight speech signals for the interference signal. The speech signals arrived from eight different directions simultaneously. The microphone signals were formed as convolutive mixtures of measured impulse responses and speech signals. The input signal-to-interference ratio (SIR) was set to 0 dB. We placed virtual microphones between two real microphones at regular intervals, and thus the interpolation parameter α of the ith virtual microphone was

$$\alpha = \frac{i}{N+1}, \tag{26}$$

where N is the number of inserted virtual microphones. The speech was enhanced using microphone arrays consisting of two real microphones and N virtual microphones, thus giving $(N+2)$ channels in total. In these

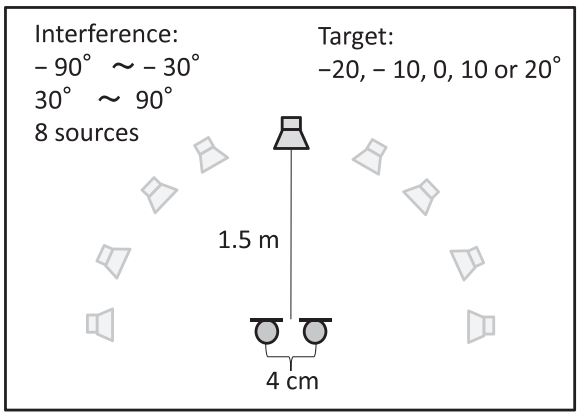

Fig. 3 Source and microphone layout in experiments

Table 1 Experimental conditions

Number of real microphones	2
Number of virtual microphones N_v	0–9
Real microphone interval	4 cm
Input SIR	0 dB
Reverberation time	640 ms
Sampling rate	8 kHz
FFT frame length	1024 samples
FFT frame shift	256 samples
Speech-enhanced period length	20 s
Target-active period length $\|\theta_T\|$	10 s
Target-inactive period length $\|\theta_I\|$	10 s

experiments, the first real microphone channel, expressed as $\alpha = 1$, was chosen as the reference for scale compensation as described in Section 4.2.

5.2 Results and discussion

To evaluate the performance of the beamformer, we used objective criteria, namely, the signal-to-distortion ratio (SDR) and SIR [18]. We show the mean SDR and SIR values for the 15 combinations of target DOA and speech samples.

Figure 4 shows the relationship between the speech enhancement performance and β for different numbers of virtual microphones, and Fig. 5 shows the relationship between the performance and the number of virtual microphones for several β values. The input SIR was set to 0 dB. This means that the target signal was larger than each of the eight interference signals. Under this condition, the time-frequency component of the target signal was able to be captured. Note that when the number of virtual microphones is zero, the beamformer is processed solely using real microphone signals. According to Fig. 4a, the SDR improved when a small number of virtual microphones are inserted for all values of β. According to Fig. 4b, the SIR improved as the number of virtual microphones increased for all values of β. The SDR decreases when a number of virtual microphones are introduced with a parameter β of around 0 as shown in Fig. 4a. However, there is a particular improvement in the SDR when β is far from 0.

Setting β at values of around 0, the interpolation may become almost linear, which could cause rank deficiency in the covariance matrices. Figure 6 shows the relationship

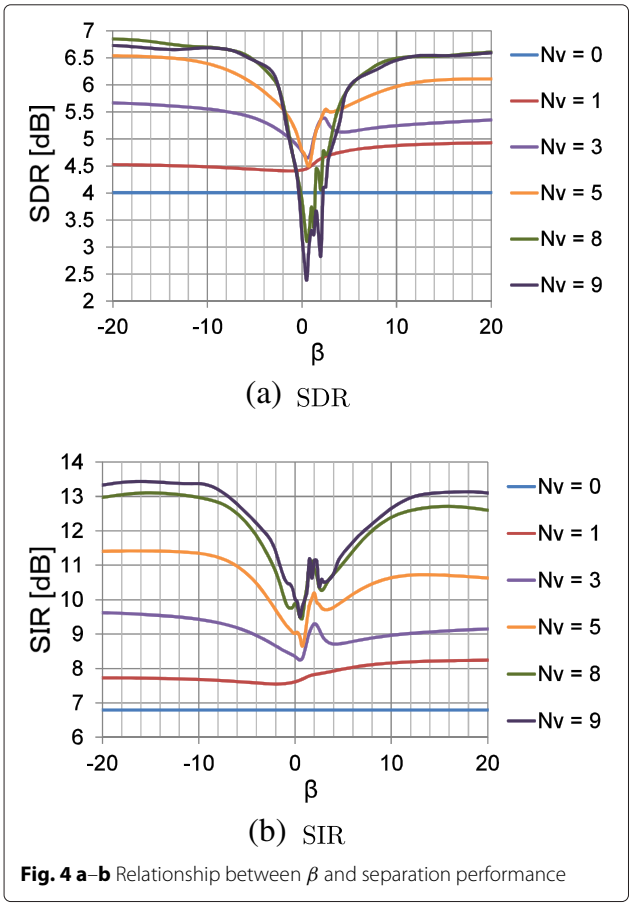

Fig. 4 a–b Relationship between β and separation performance

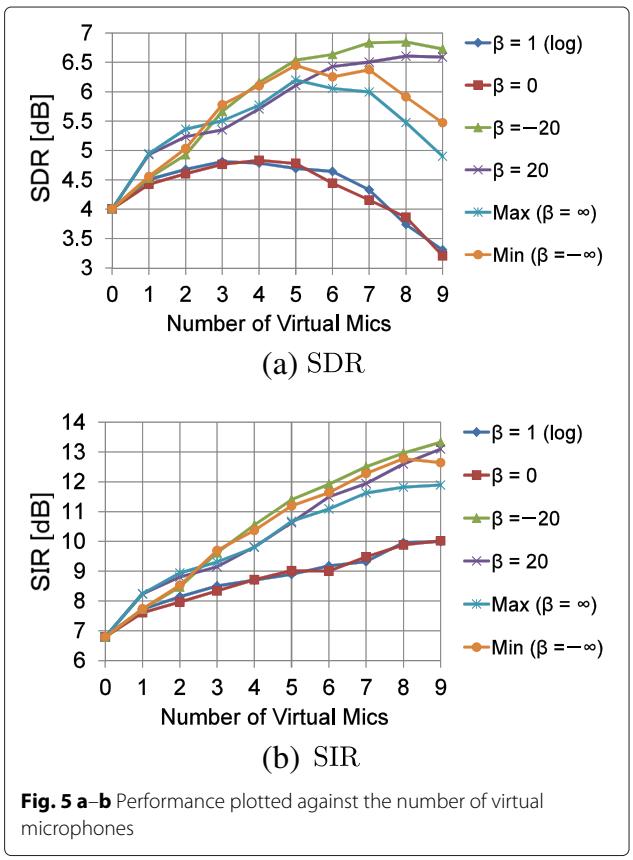

Fig. 5 a–b Performance plotted against the number of virtual microphones

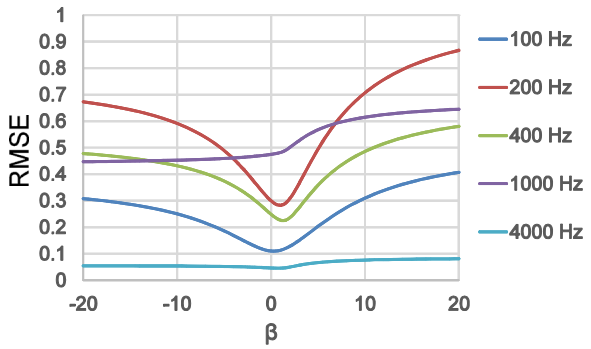

Fig. 6 Relationship between β and RMSE difference from linear interpolation

between β and the difference from linear interpolation. The difference is defined as the root-mean-square error (RMSE) between the signal generated by the proposed interpolation and the one generated by linear interpolation when the interpolation parameter α is set to 0.5, i.e., the virtual microphone is set at the midpoint between two real microphones.

$$\mathrm{RMSE}\left(\omega,\alpha\right)=\sum_{t}\left|\nu_{\beta}\left(\omega,t,\alpha\right)-\nu_{\mathrm{lin}}\left(\omega,t,\alpha\right)\right|^{2}, \quad (27)$$

$$\nu_{\mathrm{lin}}\left(\omega,t,\alpha\right)=\left(1-\alpha\right)x_{1}\left(\omega,t\right)+\alpha x_{2}\left(\omega,t\right). \quad (28)$$

A larger RMSE shows that the results of interpolation is far from that obtained by linear processing. As a virtual microphone signal is similar to a linearly generated signal, the covariance matrices tend to cause rank deficiency, resulting in the distortion of the output signal of the beamformer. According to Fig. 6, setting β to a value of approximately two, for which the β-divergence-based amplitude interpolation becomes linear, the result of interpolation of the entire signal is close to that of the linear interpolation, particularly in the 100–400-Hz frequency range. By contrast, upon setting β to a value distant from zero, the interpolation becomes far from linear. Accordingly, the nonlinearity of the interpolation is improved and the rank deficiency of the covariance matrices is reduced.

6 Conclusions

In this paper, we proposed a new array signal processing technique involving the virtual microphones to increase the number of channels virtually and to improve the performance of speech enhancement. Virtual microphone signals are generated by the interpolation of the phase and amplitude of a complex signal. The phase is interpolated linearly in accordance with a plane wave propagation model, and the amplitude is interpolated using a β-divergence-based method, that allows parameter adjustment. We also applied the virtual increase in the number of channels to speech enhancement using a maximum SNR beamformer. The speech enhancement performance using virtual microphones was evaluated in the experiments, and we investigated the relationship between the performance and the value of β. According to the experimental results, the speech enhancement performance for multiple directional sources was improved by introducing virtual microphones. By setting β to values far from 0, the performance was further improved when more virtual microphones were introduced. We also evaluated the speech enhancement performance for a mixture of target speech and real environmental noise and confirmed that there was an improvement. These results confirmed the effectiveness of the virtual microphone technique for speech enhancement.

Competing interests
The authors declare that they have no competing interests.

About the Authors
Hiroki Katahira was born in Kashima, Ibaraki, Japan. He received B.Sc. and M.Eng. degrees from University of Tsukuba in 2013 and 2015 respectively. His research interests include acoustic signal processing, specifically, microphone array signal processing. He had been engaged in research of speech enhancement with microphone array while he was in graduate school.
Nobutaka Ono received the B.E., M.S., and Ph.D degrees in Mathematical Engineering and Information Physics from the University of Tokyo, Japan, in 1996, 1998, 2001, respectively. He joined the Graduate School of Information Science and Technology, the University of Tokyo, Japan, in 2001 as a Research Associate and became a Lecturer in 2005. He moved to the National Institute of Informatics, Japan, as an Associate Professor in 2011. His research interests include acoustic signal processing, specifically, microphone array processing, source localization and separation, music signal processing, audio coding and watermarking, and optimization algorithms for them. He is the author or co-author of more than 180 articles in international journal papers and peer-reviewed conference proceedings. He was a Tutorial speaker at ISMIR 2010, a special session chair in EUSIPCO 2013 and 2015, a chair of SiSEC (Signal Separation Evaluation Campaign) evaluation committee in 2013 and 2015. He has been an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing since 2012. He has been a member of IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee since 2014. He is a senior member of the IEEE Signal Processing Society, and a member of the Acoustical Society of Japan (ASJ), the Institute of Electronics, Information and Communications Engineers (IEICE), the Information Processing Society of Japan (IPSJ), and the Society of Instrument and Control Engineers (SICE) in Japan. He received the Sato Paper Award and the Awaya Award from ASJ in 2000 and 2007, respectively, the Igarashi Award at the Sensor Symposium on Sensors, Micromachines, and Applied Systems from IEEJ in 2004, the best paper award from IEEE ISIE in 2008, Measurement Division Best Paper Award from SICE in 2013, the best paper award from IEEE IS3C in 2014, the excellent paper award from IIHMSP in 2014 and the unsupervised learning ICA pioneer award from SPIC.DSS in 2015.
Shigeki Miyabe received B.E. degree from Kobe University in 2003, and received M.E. and Ph.D. degrees from Nara Institute of Science and Technology (NAIST) in 2005 and 2007, respectively. From 2008 to 2009, he was a Visiting Scholar with Georgia Institute of Technology. In 2009, he joined the Graduate School of Information Science and Technology, University of Tokyo as a researcher, and became Assistant Professor in 2010. He is currently Assistant Professor of Life Science Center of Tsukuba Advanced Research Alliance, University of Tsukuba since 2011. He received the Awaya Prize from the Acoustic Society of Japan in 2012. He is a member of IEEE and ASJ.
Takeshi Yamada received B. Eng. degree from Osaka City University, Japan, in 1994, and M. Eng. and Dr. Eng. degrees from Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. He is presently an associate professor with Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. His research interests include speech recognition, sound scene understanding, multi-channel signal processing, media quality assessment, and e-learning. He is a member of the IEEE, the IEICE, the IPSJ, the ASJ, and the JLTA.

Shoji Makino received B. E., M. E., and Ph. D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981. He is now a Professor at University of Tsukuba. His research interests include adaptive filtering technologies, the realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech, and acoustic signal processing for speech and audio applications.
He received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2015, the IEEE Signal Processing Society Best Paper Award in 2014, the ICA Unsupervised Learning Pioneer Award in 2006, the IEEE MLSP Competition Award in 2007, the Achievement Award of the Institute of Electronics, Information, and Communication Engineers (IEICE) in 1997, and the Outstanding Technological Development Award of the Acoustical Society of Japan (ASJ) in 1995, the TELECOM System Technology Award in 2015 and 2004, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002. He is the author or co-author of more than 200 articles in journals and conference proceedings and is responsible for more than 150 patents. He was a Keynote Speaker at ICA2007, a Tutorial speaker at EMBC2013, Interspeech2011 and ICASSP2007.
He has served on IEEE SPS Technical Directions Board (2013-14), IEEE SPS Awards Board (2006-08) and IEEE SPS Conference Board (2002-04). He was a member of the IEEE Jack S. Kilby Signal Processing Medal Committee (2015-) and the James L. Flanagan Speech and Audio Processing Award Committee (2008-11). He was an Associate Editor of the IEEE Transactions on Speech and Audio Processing (2002-05) and an Associate Editor of the EURASIP Journal on Advances in Signal Processing (2005-12). He was the Chair of SPS Audio and Acoustic Signal Processing Technical Committee (2013-14) and the Chair of the Blind Signal Processing Technical Committee of the IEEE Circuits and Systems Society (2009-10). He was the General Chair of WASPAA2007, the General Chair of IWAENC2003, the Organizing Chair of ICA2003, and is the designated Plenary Chair of ICASSP2012.
Dr. Makino is an IEEE SPS Distinguished Lecturer (2009-10), an IEEE Fellow, an IEICE Fellow, a Board member of the ASJ, and a member of EURASIP.

Acknowledgements
This work was supported by the National Institute of Informatics (grant no. 2013-5).

Author details
[1]University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan. [2]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan. [3]SOKENDAI (The Graduate University for Advanced Studies), Hayama, Japan.

Received: 28 April 2015 Accepted: 30 December 2015
Published online: 29 January 2016

References

1. M Brandstein, D Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. (Springer, New York, 2001)
2. S Makino, T-W Lee, H Sawada, *Blind Speech Separation*. (Springer, New York, 2007)
3. Y Izumi, N Ono, H Sagayama, Sparseness-based 2ch BSS using the EM algorithm in reverberant environment. Proc. WASPAA, 147–150 (2007)
4. H Sawada, S Araki, S Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. on Audio, Speech & Language Processing. **19**(3), 516–627 (2011)
5. M Nakano, H Kameoka, J Le Roux, Y Kitano, N Ono, S Sagayama, Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence. Proc. MLSP, 283–288 (2010)
6. H Saruwatari, S Kajita, K Takeda, F Itakura, Speech enhancement using nonlinear microphone array based on complementary beamforming. IEICE Trans. Fundamentals. **E82–A**(8), 1501–1510 (1999)
7. S Miyabe, B-HF Juang, H Saruwatari, K Shikano, Analytical solution of nonlinear microphone array based on complementary beamforming. Proc. IWAENC, 1–4 (2008)
8. P Chevalier, A Ferréol, L Albera, High-resolution direction finding from higher order statistics: The 2q-MUSIC algorithm. IEEE Trans. on Signal Processing. **53**(4), 2986–2997 (2006)
9. Y Sugimoto, S Miyabe, T Yamada, S Makino, B-HF Juang, Employing moments of multiple high orders for high-resolution underdetermined DOA estimation based on MUSIC. Proc. WASPAA, 1–4 (2013)
10. Y Hioka, T Betlehem, Under-determined source separation based on power spectral density estimated using cylindrical mode beamforming. Proc. WASPAA, 1–4 (2013)
11. G Del Galdo, O Thiergart, T Weller, EAP Habets, Generating virtual microphone signals using geometrical information gathered by distributed arrays. Proc. HSCMA, 185–190 (2011)
12. K Kowalczyk, A Craciun, EAP Habets, Generating virtual microphone signals in noisy environments. Proc. EUSIPCO, 1–5 (2013)
13. S Rickard, O Yilmaz, On the approximate W-disjoint orthogonality of speech. Proc. ICASSP. **1**, 529–532 (2002)
14. O Yilmaz, S Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. on Signal Processing. **52**(7), 1830–1847 (2004)
15. C Févotte, J Idier, Algorithms for nonnegative matrix factorization with the β-divergence. Neural Comput. **23**(9), 2421–2456 (2011)
16. HL Van Trees, *Optinum Array Processing*. (Wiley, New York, 2002)
17. S Araki, H Sawada, S Makino, Blind speech separation in a meeting situation with maximum SNR beamformers. Proc. ICASSP. **I**, 41–45 (2007)
18. E Vincent, R Gribonval, C Févotte, Performance measurement in blind audio source separation. IEEE Trans. on Audio, Speech & Language Processing. **14**(4), 1462–1469 (2006)